ChinaXiv合作期刊
图书情报工作
第63卷 第4期 2019年2月

用于科学结构分析的混合网络社团划分方法述评*

■ 张瑞红[1,2] 陈云伟[1,2] 邓勇[1,2]

[1] 中国科学院成都文献情报中心 科学计量与科技评价研究中心(SERC) 成都 610041

[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘要：[**目的/意义**]复杂网络的社团结构研究已逐渐成为科学家借助文献数据开展科学结构研究的有力工具,社团划分效果的不同对科学结构的解读有着举足轻重的影响。本文对混合网络社团划分方法进行梳理,以期对该领域的相关研究提供借鉴参考。[**方法/过程**]通过文献调研,阐明混合网络的概念与类型,从网络构建或算法革新角度对各类型混合网络的社团划分研究进行概述,也对支撑混合网络社团划分的经典算法进行简介。[**结果/结论**]通过系统地梳理总结不同类型混合网络的社团划分工作,为后续的网络分析研究提供研究的视角和方法,同时揭示其在科学结构研究中所面临的挑战与所具有的现实意义,展望今后可能进一步拓展的相关研究方向。

关键词：混合网络 社团划分 聚类分析 合作 引用

分类号：G250

DOI:10.13266/j.issn.0252-3116.2019.04.016

科学研究的日益复杂性与交叉性使学科边界变得模糊,进而使科学结构越来越难以被清晰地认识。科学结构是长期形成的、固有的、不以人们意志为转移的客观存在[1],是科学内在逻辑的外在体现,反映在科学的门类结构、科学的学科结构、科学的知识结构上[2]。虽然科学的内在本质是客观不变的,但其外在体现却随着人类对科学认知的加深而不断演化。如何有效地发现科学结构已成为知识发现研究的焦点问题,对探索学科演化、发现学科交叉渗透、挖掘前沿方向具有重要价值。2002年社团(community)概念被正式提出后,社团划分研究逐渐受到关注,而社团划分问题本质上是关联节点间的聚类问题。近年来,基于文献网络(如合作网络、引文网络等)的社团结构研究已成为科学家借助文献数据开展科学结构研究的有力工具。科学家合作网络的结构、引文网络的结构等在一定程度上反映的正是科学的学科结构或知识结构。

例如,2002年M. Girvan和M. E. J. Newman首次提出社团概念时,就利用GN算法对圣塔菲研究所1999-2000年间科学家的合作网络的主成分(118位科学家)开展了社团划分研究(见图1)[3],将这些科学家分成了4个社团(基于代理的模型研究经济和交通问题、生态学的数学模型、统计物理、RNA结构)。随后,有关合作网络的社团研究大量涌现。R. Lambiotte和P. Panzarasa在2009年通过对合作网络进行社团划分,研究了科学合作模式是如何促进知识创造和扩散的[4];L. A. Moliner等在2017年研究了人才管理领域科学家合作网络社团的演化历程[5],丰富了人才管理动力学的相关研究,提供了关于研究人员之间合作原因与合作模式的证据;J. Zheng等在2017年基于单本期刊的作者共著网络开展了社团的演化研究[6],发现了分析合作者社团演化更有效的综合指数与生命周期策略,为通过合作网络来动态观察学术共同体的演化研究提供了新思路。同时,引文网络分析的内涵和方法也随着社会网络分析方法的发展得以不断丰富[7],通过引文网络社团分析能够更准确地揭示科学结构和发展过程[8]。例如,Y. Kajikawa[9]等利用FN算法通过分析引文网络社团随时间的变化情况来识别新型研究领域,以拓展学科的知识结构;陈云伟提出了一种基于样本加权的引文网络社团划分方法,以Louvain社团划分方法为算法基础,将科学论文用向量空间模型表示,利

* 本文系国家重点研发计划现代服务业重点专项"专业内容知识服务众智平台与应用示范"(项目编号:2017YFB1402400)研究成果之一。

作者简介: 张瑞红(ORCID:0000-0001-5786-4182),硕士研究生;陈云伟(ORCID:0000-0002-6597-7416),研究员,硕士生导师;邓勇(ORCID:0000-0001-9179-0500),研究员,硕士生导师,通讯作者,E-mail:dengy@clas.ac.cn。

收稿日期:2018-05-21 修回日期:2018-09-26 本文起止页码:135-141 本文责任编辑:王传清

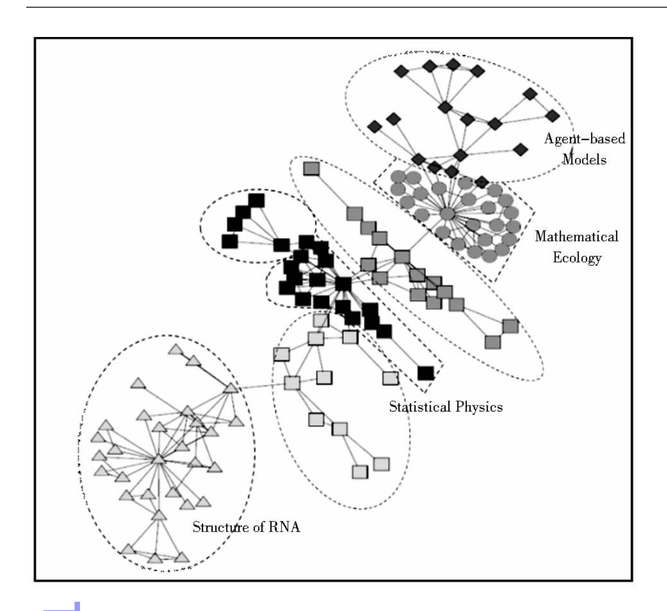


图 1 圣塔菲研究所 1999 – 2000 年间 271 位科学家的合作网络(118 人)社团划分[3]

用余弦相似度方法计算相邻论文之间的相似度并作为引用边的权重。然后又进一步综合考虑节点结构与内容相似度对网络进行重构,社团划分结果十分明显[10]。近年来具有代表性的工作是莱顿大学 L. Waltman 和 N. J. V. Eck 在 2013 年开发的 CitNetExplorer 软件,集成了 SLM 社团划分算法以用于引文网络的社团划分研究[11]。

然而,尽管图书情报领域的研究人员对合著网络、引文网络等开展了大量的社团研究工作,但多数研究针对的仅是单一节点类型的网络(如仅以作者或论文为节点的同构网络),存在分析对象单一、关联关系单一,对科学结构的揭示不精细、不完整等不足。为此,近年来有研究人员开始研究对具有多种节点类型或多种关系的网络进行社团划分,以期提升科学结构分析的效果。

网络的类型对社团划分的效果及要揭示的科学结构会有一定影响,特别是混合网络中不同类型的对象及其之间不同类型的相互关系在揭示网络所携带的丰富语义方面存在更丰富的功能,也可能得到多种不同的挖掘结果[12]。因此,本文以混合网络社团划分研究为出发点,检索图书情报领域国内外相关研究文献,对近年来有关混合网络社团划分方法的研究进行梳理和述评。鉴于混合网络的社团划分过程包含网络构建和社团划分算法两个核心工作,而社团划分算法的发展相对较为成熟,当前针对混合网络社团划分方法的研究工作,核心贡献多数都是针对如何构建有效的混合网络方面,在社团实现方面多数采用已有的成熟算法,本文重点评述混合网络的构建方法,讨论不同类型混合网络的社团划分效果,以期揭示利用混合网络划分方法研究科学结构及其演化规律功能的发展脉络,以及今后相关研究可能的发展趋势。同时,具有多节点类型多关系网络的社团划分算法更具难度与挑战性,因此许多研究工作集中在社团划分算法的推演上,此类研究也在下文做简要概述。最后,对当前图书情报领域常用的社团划分算法进行简单总结,供有兴趣开展相关研究工作的学者参考。

1 混合网络的概念与类型

本文定义的"混合网络"是指含有多种节点类型或多种关系的网络,即网络中同时包含作者和论文两种或两种以上类型的节点,或网络的边涵盖了合作、引用或主题相似等两种或两种以上的关系。根据韩家炜和 B. Taskar 等对网络的定义,混合网络本质上属于异构网络的范畴,即多种类型节点与多种关系的边所组成的网络是异构的[12-13]。但鉴于异构网络的概念强调的是网络结构层面的复杂性,而本文所分析的多节点、多关系集成在一起所形成的网络强调的是功能的丰富性,故而提出"混合网络"的概念,便于图书情报研究人员将研究焦点聚焦到功能提升上,而不是把网络变得更复杂。

通过与图书情报领域针对单一节点类型网络(如引文网络、合作网络)社团划分研究的工作进行比较,发现按照网络节点与边的类型可以将混合网络分为三类:第一类,单类型节点多关系网络,如以作者为单一节点的网络,同时包含合作和引用两种关系;第二类,多类型节点多关系网络,如网络中同时包含作者和论文两种节点,同时包含合作和引用两种关系;第三类,多类型节点单关系网络,如网络中同时包含作者和论文两种节点,但仅有引用一种关系。

下文主要从这三种类型的混合网络出发,分别对其社团划分的相关工作进行阐述和分析。

2 单类型节点多关系混合网络的社团划分

在单类型节点多关系网络中,可以通过节点间多种不同的关系对网络边的含义赋予更丰富的内涵,再进行聚类或社团划分。当前对选择哪些不同的关系进行结合有两种不同的方向:其一是基于研究目的,将节点间不同类型的关系直接叠加在网络中,即多关系组合(relation combination),其二是将多种关系融合成一种新的关系后再分析研究对象的关系特征等,即多关系融合(relation fusion)。

2.1 多关系组合方法

将多关系组合方法应用于学科领域的科学结构分析中,主要包括:①引用关系与共词关系的组合。最具代表性的工作是 H. Small 在 1998 年将引用和共词两种关系组合在一起来揭示文献间的直接连接关系和间接连接关系,进而作为一个涉及分层聚类、聚类的排序以及公共坐标投射方法的框架,支撑科学结构地图的可视化呈现研究[14]。其他工作还包括:C. Calero-Medina 等在 2008 年利用共词和引用关系组合的方法确定了那些影响某领域一段时间的文章,通过将这些文章与某领域早期具有影响力的传统研究联系起来,分析了科学出版物间知识的创造和流动过程,对后续利用多种方法结合的相关研究具有启发作用[15];侯跃芳等在 2007 年应用内容词与引文共引聚类分析,既揭示了妊娠糖尿病专题研究的发展现状又验证了聚类效果,为组合分析方法应用到专题研究开了先河[16];张晗等在 2007 年利用共词分析与文献的引用次数相结合的方法,基于 PubMed 数据库,全面探索了消化性溃疡领域学科领域的发展进程[17],验证了主题词共词分析与主题的被引频次相结合更易于检测学科热点。②合作关系与引用(含同被引、文献耦合)关系的组合。例如,K. Larsen 等在 2008 年将合著和同被引两种关系组合起来用于测度太阳能电池研究知识网络的中心点,得出了区分新研究领域发展的早期和晚期阶段的重要性,以及要在科技领域对学习过程和知识传播开展系统性观察[18];陈伟等在 2014 年以我国“985”高校为节点构建了合著网络和被引网络,对两种网络的基本结构特征、网络关联性质、社团特征和重要节点进行了联合分析,揭示了“985”高校科研合作网络的复杂性特征和发展趋势[19],为研究高校间合作与引用打开了新视角。

这些研究均对两种或两种以上的关系进行了组合使用,可以从不同角度更全面挖掘出研究对象的特征,更有效地揭示了科学结构及演化问题。然而,可选的组合很多,为了判断如何进行有效的组合能实现最佳的效果,E. Yan 等[20]对图书情报领域经常分析的合作网络、主题网络、引用网络等进行了相似度测量,发现主题网络与合作网络具有最低的相似度,共引网络与引文网络具有较高的相似度,文献耦合网络与共引网络也具有较高的相似度,共词网络与主题网络依然具有较高的相似度。研究中对具有较高相似度的共引网络与引文网络进行组合,发现因为网络相似度较高使分析结果类似,对于问题的全面分析没有太大帮助。因此,关系组合应首先从基于引用与非引用、基于社交与认知这两个维度入手,对相似度较低的网络进行组合以揭示更多信息。

2.2 多关系融合方法

与多关系组合方法不同,多关系融合方法是对多种关系进行融合处理,该方法源于对网页的聚类或分类研究。按照融合阶段的不同,可分为两种类型:一种是社团合并,即分别将不同数据源进行聚类,再通过一定的算法将不同的聚类结果合并到新的聚类。另一种是核融合,即将多源数据的相似度矩阵或距离矩阵整合为一个新的独立矩阵,再用相关算法进行聚类或其他多元统计分析。

在单类型节点多关系混合网络的聚类合并研究方面,X. X. Yin 等在 2015 年提出一种叫做 CROSSCLUS 的简单半监督方法,该方法根据用户选择的一组与聚类目标相关的特征,对多关系的对象进行多次聚类评估[21];L. Wei 等在 2015 年针对多关系的数据使用相关分析方法,将不同聚类之间的距离计算为每个聚类中心点的距离,并为之赋权重,保证了实体之间聚类的效率与聚类的精度[22];丁志军等在 2017 年提出分部多关系聚类方法,是聚类集成关系融合的典型研究。该方法根据实体间的不同关系对实体进行聚类,再根据聚类结果对不同关系的重要性进行加权赋值,最后整合为单关系网络再进行聚类,该方法经过多组公开数据集的实验,证明其可以有效地提升聚类精度[23]。以上聚类方法在效率与精度上均有所提升,对利用聚类方法展示科学结构的研究提供了更可靠、更准确的方法基础。

在单类型节点多关系混合网络的核融合研究方面,近年来有代表性的相关工作是综合考虑学术论文的文本属性及链接属性的混合聚类方法,如围绕 W. Glanzel 提出的综合引文耦合和文本相似度的“引文－文本”混合聚类算法[24－25]的一系列相关研究,证明了混合聚类方法比使用单一的方法进行社团划分的准确率更高。首先,W. Glanzel 等借鉴网页内容与链接分析相结的思想,将文献间基于词的关系与基于文献耦合的关系结合到一起,研究结果证明了这种方法在揭示研究领域结构上的有效性[26];其次,张琳等使用基于期刊交叉引用的聚类算法来验证和改进基于期刊的学科分类方案[27];再者,W. Glanzel 团队还将期刊的交叉引用同文本挖掘进行整合,验证并提高了现有的主题分类方案[28]。此外,王小梅在近年来陆续发布的系列《科学结构地图》中,也是采用的 W. Glanzel 团队的混合聚类方法。

在关系融合研究上,W. Glanzel 团队的研究主要集

中在对引用关系与文本这两种互为补充关系的信息挖掘上,并没有涉足其他两两独立关系的研究。如代表基于引用关系的引文网络与基于社交认知的合作网络之间的混合聚类效果如何,这是今后需要进一步研究与探索的。

3 多类型节点多关系混合网络的社团划分

多类型节点多关系的网络是相较于传统的网络而定义的,即网络中若存在多种实体类型与多种关联关系,可以视为异构信息网络。在图书情报领域,文献信息网络就是一种具有多种实体类型与关系的异构信息网络,主要涉及文章、期刊、作者和关键词 4 类实体。其中,文章与期刊、文章与作者、文章与关键词都具有关系。因为信息在异构节点与关系间的流动不同于同构网络,很多基于同构网络的分析方法不适用于异构信息网络,所以对诸如此类的网络聚类或社团划分研究多集中于对算法的推新与改进上。目前,研究多类型节点多关系网络的社团划分主要有三种思路,分别是基于排序的方法、基于元路径的方法以及异构网络同构方法。

3.1 基于排序的方法

将排序方法应用于社团划分或聚类中,排序与聚类可以相辅相成。最先基于异构信息网络的排序聚类算法是 RankClus[29],其原理是对网络中的不同节点不断地进行聚类与排序,直到研究对象的聚类明晰化;之后出现了许多相似的排序聚类算法,如 NetClus 算法[30]、ENetClus 算法[31]、ComClus 算法[32]等,其中 NetClus 算法主要是针对星型结构的网络,该算法可以高效地产生聚类结果与排名结果。赵焕对 NetClus 算法进行改进,提出基于异构网络的 MAO-Netclus 算法,对 Web 服务系统的三种对象所构成的多类型节点多关系网络进行聚类分析,实现对 Web 服务推荐的改进[33];童浩等提出一种针对异构信息网络的基于排名与协同聚类的 RankCoClus 算法,实验结果显示该方法的聚类性能更优越[34]。

3.2 基于元路径的方法

基于元路径的方法是针对链接关系的方法,网络中的不同链接传递着不同的信息,对聚类的效果具有一定的影响,而异构网络中的不同链接路径构成了不同的元路径。代表性的方法是 Y. Sun 等在 2011 年提出来的 PathSim 方法[35],该方法是一种基于元路径的相似性度量方法。由于该方法只是针对同类节点计算相似度,因此随后又出现了针对非同类节点的相似度度量方法,算法性能不断在提升。同时,基于元路径的聚类方法也相继涌现,其中 PathSelClus 方法[36]研究了不同元路径对节点聚类效果的影响,该方法在元路径选择等方面需要较强的假设条件;而 GenClus 算法[37]是一种考虑链接关系强度的聚类方法,通过用户指导,确定节点属性与链接关系,并能够自动学习以构建不同的链接强度,使聚类效果得到改善;李立基于元路径的方法提出了一种启发式的搜索与剪枝策略,有效地选择出与用户指导信息一致的路径并避免了宽度优先遍历搜索的信息缺失问题。在此基础上,李立对同构网络的社团划分算法进行拓展,提出将关系抽取与元路径加权相结合的社团划分框架,并在真实数据集上验证了该方法的有效性与准确性[38];王锐在其研究中也提出了一种考虑权重的元路径社团划分算法 HCD,不仅有效地划分出多条元路径的社团,而且可以探测出重叠社团[39]。

3.3 异构网络同构方法

同构网络的社团划分算法相对成熟,因此将异构网络降维重构为同构网络也是一种可行的方法。目前,针对异构网络的降维方法主要有线性降维分析(linear discriminant analysis, LDA)、主成分分析(principal component analysis, PCA)、非负矩阵分解(non-negative matrix factorization, NMF)以及主题模型(topic model)等。重构方法主要是将异构网络重构为二分图的方法。基于以上方法,王婷在 2016 年提出一种高效快速的异构网络社团探测算法[40],首先对异构社交网络数据进行降维,然后将异构网络重构为二分图,为了在社团划分中不使信息丢失,利用标签传播的方法进行社团划分[37],该方法具有一般性,可以推广应用到许多实际场景中。

典型的异构信息网络聚类算法是基于元路径的方法与基于排序的方法,前者较后者而言,省去了繁琐的排序迭代过程,但是却需要用户先验经验的指导,各有利弊。异构网络同构方法便于理解,但过程较为复杂。若要观察领域复杂的异构网络的科学结构,在异构网络的前期处理中,应用基于排序的方法、基于元路径的方法或异构网络同构的方法是必要的。

4 多类型节点单关系混合网络的社团划分

多类型节点单关系网络的特点是节点类型呈现多样性,关于该类网络的社团划分研究鲜少,但可以从网络构建角度为社团划分提供前期工作的参考。若对该类网络进行社团划分或聚类,需要理解节点类型的含

义。

4.1 多重属性的节点

有些网络中的多类型节点本质是实体多重属性的体现。比如合作网络中,节点一般是作者或研究人员,并不区分其社会属性。但是严格来说,研究人员是有多重属性的,包括文章属性(关键词、主题等)、特征属性(年龄、职称等)以及社会属性(学生、教师)。王炎等利用专家学者的不同属性对专家学者学术网络进行了理论与方法的探究,基于多元数据构建了专著专家合作网络、专家主题网络、专利专家合作网络等,更加准确地刻画了专家间的显性与隐性合作网络[41];雷雪等根据作者贡献度,将文章合作者区分为第一作者与其他作者,构建了基于两类节点类型的有向合作网络,并与传统无向合作网络进行对比,以探索更有效的科研分析方法[42];谭宗颖等在研究国际合作时,将国家区分为主导国家和其他国家,构建了以中国为主导的有向国际合作网络,并进行主题内容分析[43]。

4.2 多种实体的节点

有些网络中的多类型节点本质上是不同实体的体现。王朋等在研究校企合作网络时,构建了科研人员与纳米类专利之间的关系网络,揭示了以清华大学为主的产学研纳米技术合作网络的拓扑结构[44];马艳艳等进一步拓展研究对象,利用中国大学与企业的专利申请数据描绘了高校与企业专利申请合作网络图,并进行了网络特性的分析,发现中国的产学研合作具有很大上升空间[45]。

不难发现,多类型节点单关系的研究主要集中在合作网络上。目前对构建多类型节点单关系网络并进行社团划分的研究工作相对较少,但是该类网络的社团划分或聚类更有利于对科学结构形成过程中的继承、从属关系进行清晰判断,也是今后可能的研究聚焦点。

5 支撑混合网络社团划分的算法简介

社团划分方法是研究复杂网络结构的重要方法,2002 年 M. Girvan 和 M. E. J. Newman 提出一种分裂算法－GN 算法[3],开启了社团研究的热潮。GN 算法是通过不断移除介数最高的边而实现社团划分的;从另一个划分角度,M. E. J. Newman 又提出一种基于聚合的贪婪算法[46],即将网络中的每个节点都作为一个独立的团簇,在划分过程中节点不断地进行合并形成社团。随后为了衡量社团划分结果的好坏, M. E. J. Newman 和 M. Girvan 于 2004 年提出模块度函数 Q[47],一般认为,Q 值越大,社团划分越好。为了解决大型网络社团发现效率偏低的问题,V. Blondel 等[48]于 2008 年提出 Louvain 社团划分算法,R. Rotta 和 A. Noack 在 2011 年对 Louvain 算法进行了优化,提出了 Louvain 算法的多级细分[49]。在此基础上,L. Waltman 和 N. J. V. Eck 在 2013 年改良提出 SLM 算法,SLM 的特点在于允许已经被划分社团的点重新进行社团划分[11]。

关于这些社团划分算法的详细介绍和比较研究,可以参考时京晶[50]、陈云伟和张瑞红[51]等研究成果。

6 讨论与展望

本文梳理了用于科学结构分析的混合网络社团划分方法在图书情报领域的最新研究进展,发现对于单类型节点多关系的混合网络,有两种方式来对多关系进行处理,分别是多关系组合与多关系融合。多关系组合较为简单,选择两种或两种以上的分析方法即可实现关系组合方法对问题的解决。在选择分析方法时,要有所依据,需要对不同组合效果进行科学评估,最好选择不同维度的分析方法。多关系融合方法主要集中在混合网络构建或社团划分算法的革新改进上。然而,关系的融合是比较复杂的工作,选择哪些关系进行融合以及融合效果的判定,都还需要开展研究进行探索。

对于更为复杂的多类型节点多关系网络,由于节点属性的多样性与关系的复杂性,当前的研究工作相对较少。研究重点和难点包括探究多类型节点多关系网络信息挖掘的原理与方法、如何准确构建模拟现实世界的模型以及对多节点或多关系重要性的判别等。其研究前沿不仅局限于对网络构建方法与社团划分或聚类的探索上,还有信息扩散、语义搜索、智能查询等。由于挖掘异构信息网络的难度较大,因此该类研究更具挑战性与现实价值,也是今后信息网络研究的重要方向之一。

在揭示科学结构方面,单一网络的社团划分研究已经相对成熟,而混合网络的社团划分研究正处于成长阶段。对于打破了传统单一网络研究局限的混合网络而言,为后续的网络分析研究提供了新的视角和方法,并且可以挖掘出隐藏在实体间不同链接间的丰富信息,在理论和实践上都是一次全新的提升和尝试。同时,混合网络的社团划分在分析科学结构、描述知识发展以及分析学科交叉等方面仍然有许多值得探索的问题。科学研究是一个复杂的系统,本文讨论的数据基础都是文献,这仅仅是科研产出的一部分。科学研究还涉及到科技战略、规划、项目、资助等大量的信息,这些信息也都是与科学结构密切相关的。未来的研究

中,还可以拓展数据基础,从更加全面的角度,利用丰富的数据类型和关系类型,充分理解和揭示科学结构。

参考文献:

[1] 卫军朝,蔚海燕. 科学结构及演化分析方法研究综述[J]. 图书与情报,2011(4):48-52.

[2] 赵红洲. 论科学结构[J]. 中州学刊,1981(3):59-65,133.

[3] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. PNAS, 2002, 99(12): 7821-7826.

[4] LAMBIOTTE R, PANZARASA P. Communities, knowledge creation, and information diffusion [J]. Journal of informetrics, 2009, 3(3): 180-190.

[5] MOLINER L A, GALLARDO-GALLARDO E, PUELLES P G D. Understanding scientific communities: a social network approach to collaborations in talent management research[J]. Scientometrics, 2017, 113(3): 1439-1462.

[6] ZHENG J, GONG J Y, LI R, et al. Community evolution analysis based on co-author network: a case study of academic communities of the journal of "Annals of the Association of American Geographers"[J]. Scientometrics, 2017, 113(2): 845-865.

[7] 陈云伟. 引文网络演化研究进展分析[J]. 情报科学,2016,34(8):171-176.

[8] 邱均平, 董克. 引文网络中文献深度聚合方法与实证研究——以 WOS 数据库中 XML 研究论文为例[J]. 中国图书馆学报, 2013, 39(2):111-120.

[9] KAJIKAWA Y, YOSHIKAWA J, TAKEDA Y, et al. Tracking emerging technologies in energy research: toward a roadmap for sustainable energy[J]. Technological forecasting and social change, 2008, 75(6):771-782.

[10] CHEN Y W, XIAO X, DENG Y, et al. A weighted method for citation network community detection[EB/OL]. [2018-05-20]. http://ir.csdl.ac.cn/handle/12502/9556? mode = full&submit_simple = Show + full + item + record.

[11] WALTMAN L, ECK N J V. A smart local moving algorithm for large-scale modularity-based community detection[J]. European physical journal b, 2013, 86(11):471.

[12] 孙艺洲,韩家炜. 异构信息网络挖掘:原理与方法[M]. 北京:机械工业出版社,2016.

[13] TASKAR B, ABBEEL P, KOLLER D. Discriminative probabilistic models for relational data[J]. Algorithmic bioproceses lncs, 2002, 7(7):485-492.

[14] SMALL H. A general framework for creating general largescale maps of science in two or three dimensions: the sciviz system[J]. Scientometrics, 1998,41(1/2):125-133.

[15] CALERO-MEDINA C, NOYONS E C M. Combining mapping and citation network analysis for a better understanding of the scientific development: the case of the absorptive capacity field[J]. Journal of informetrics, 2008,2(4):272-279.

[16] 侯跃芳,崔雷,吴迪. 应用引文共引聚类-内容词分析法对学科发展的研究[J]. 情报学报,2007, 26(2):309-314.

[17] 张晗,王晓瑜,崔雷. 共词分析法与文献被引次数结合研究专题领域的发展态势[J]. 情报理论与实践,2007,30(3):378-380,426.

[18] LARSEN K. Knowledge network hubs and measures of research impact, science structure, and publication output in nanostructured solar cell research [J]. Scientometrics, 2008, 74 (1): 123-142.

[19] 陈伟,周文,郎益夫,等. 基于合著网络和被引网络的科研合作网络分析[J]. 情报理论与实践,2014,37(10):54-59.

[20] YAN E, DIING Y. Scholarly network similarities: how bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other[J]. Journal of the American Society for Information Science and Technology,2012,63 (7):1313-1326.

[21] YIN X X, HAN J, YU P S. CrossClus: user-guided multi-relational clustering[J]. Data mining and knowledge discovery,2007,15(3):321-348.

[22] WEI L, LI Y. Multi-relational clustering based on relational distance[C]//Web information system and application conference. Jinan:IEEE,2016: 297-300.

[23] 曾严昱,丁志军. 考虑重要性赋权的分部多关系聚类方法[J]. 小型微型计算机系统,2017,38(6):1227-1230.

[24] GLENISSON P, GLANZEL W, JANSSENS F, et al. Combining full text and bibliometric information in mapping scientific disciplines[J]. Information processing & management, 2005,41(6): 1548-1572.

[25] JANSSENS F, GLANZEL W, DE MOOR B. A hybrid mapping of information science[J]. Scientometrics, 2008,75(3): 607-631.

[26] JANSSENS F. Clustering of scientific fields by integrating text mining and bibliometrics [EB/OL]. [2017-12-09]. https://repository.libis.kuleuven.be/dspace/bitstream/1979/847/5/PhD_Frizo_Janssens.html.

[27] ZHANG L, JANSSENS F, LIANG L M, et al. Journal cross-citation analysis for validation and improvement of journal-based subject classification in bibliometric research [J]. Scientometrics, 2010, 82 (3): 687-706.

[28] JANSSENS F, ZHANG L, MOOR B. D, et al. Hybrid clustering for validation and improvement of subject-classification schemes [J]. Information processing & management, 2009,45(6): 683-702.

[29] SUN Y, HAN J, ZHAO P, et al. RankClus: integrating clustering with ranking for heterogeneous information network analysis[C]//Proceedings of the 12th international conference on extending database technology: advances in database technology. Petersburg: ACM, 2009:565-576.

[30] BARALIS E, BIANCO A, CERQUITELLI T, et al. NetCluster: a clustering-based framework for internet tomography[C]// IEEE international conference on communications. Dresden: IEEE,2009:1-5.

[31] GUPTA M, AGGARWAL C C, HAN J, et al. Evolutionary cluste-

ring and analysis of bibliographic networks[EB/OL]. [2018－05－20]. http://charuaggarwal.net/asonam-cluster.pdf.

[32] WANG R, SHI C, YU P S, et al. Integrating clustering and ranking on hybrid heterogeneous information network[C]//Pacific-Asia conference on knowledge discovery and data mining. Gold Coast: Springer, 2013:583－594.

[33] 赵焕. 基于异构网络聚类的 Web 服务推荐系统研究[D]. 重庆:重庆大学,2015.

[34] 童浩,余春艳. 基于排名分布的异构信息网络协同聚类算法[J]. 小型微型计算机系统,2014,35(11):2445－2449.

[35] SUN Y, HAN J, YAN X, et al. Pathsim: meta path-based top-k similarity search in heterogeneous information networks[J]. Proceedings of the VLDB endowment, 2011, 4(11):992－1003.

[36] SUN Y, NORICK B, HAN J, et al. Pathselclus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks[J]. ACM transactions on knowledge discovery from data (TKDD), 2013, 7(3):1－23.

[37] SUN Y, AGGARWAL C C, HAN J. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes[J]. Proceedings of the VLDB endowment, 2012, 5(5):394－405.

[38] 李立. 基于元路径选择和融合的异构信息网络社区挖掘算法研究[D]. 西安:西安电子科技大学,2014.

[39] 王锐. 异质信息网络中的社团发现研究与实现[D]. 北京:北京邮电大学,2017.

[40] 王婷. 异构社交网络中社区发现算法研究[D]. 北京:中国矿业大学(北京),2016.

[41] 王炎,魏瑞斌. 基于多数据源的专家学术网络构建研究[J]. 情报杂志,2016,35(12):121－126,138.

[42] 雷雪,王立学,曾建勋. 作者合著有向网络构建与分析[J]. 图书情报工作,2015,59(5):94－99.

[43] 鲁晶晶,谭宗颖,刘小玲,等. 国际合作中国家主导合作研究的网络构建与分析[J]. 情报杂志,2015,34(12):60－66,100.

[44] 王朋,张旭,赵蕴华,等. 校企科研合作复杂网络及其分析[J]. 情报理论与实践,2010,33(6):89－93.

[45] 马艳艳,刘凤朝,孙玉涛. 中国大学－企业专利申请合作网络研究[J]. 科学学研究,2011,29(3):390－395,332.

[46] NEWMAN M E. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6):066133.

[47] NEWMAN M E, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2):026113.

[48] BLONDEL V, GUILLAUME J, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008(10):155－168.

[49] ROTTA R, NOACK A. Multilevel local search algorithms for modularity clustering[J]. Journal of experimental algorithmics, 2011, 16(2): Article No. 2.3.

[50] 时京晶. 三种经典复杂网络社区结构划分算法研究[J]. 电脑与信息技术, 2011, 19(4):32－43, 79.

[51] 陈云伟,张瑞红. 用于情报挖掘的典型网络社团划分算法比较研究[J]. 数据分析与知识发现,2018,2(10):84－94.

作者贡献说明:

张瑞红:收集整理资料,撰写并修改论文;

陈云伟:提出综述撰写思路及修改意见;

邓勇:提出文章框架与撰写细节的修改意见。

A Review of Community Discovery in Hybrid Network for Science Structure Analysis

Zhang Ruihong[1,2] Chen Yunwei[1,2] Deng Yong[1,2]

[1] Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu 610041

[2] Department of Library, Information and Archives Managment, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] The study of community structure of complex networks has gradually become a powerful tool for scientists to carry out scientific structure research with literature data. In addition, the different results of community discovery play an important role in the interpretation of scientific structure. Therefore, this paper sorts out the methods of community discovery in hybrid networks, in order to provide reference and expand the ideas for the relevant researchers in the field. [**Method/process**] Through literature research, this paper mainly clarifies the concept and types of hybrid networks, and summarizes the research on community discovery of various types of hybrid networks from the perspective of network construction or algorithm innovation. Furthermore, the classical algorithm for supporting hybrid networks community discovery is also introduced. [**Result/conclusion**] Through the systematic review of the community discovery of different types of hybrid networks, it provides a new perspective and method for subsequent network analysis research, meanwhile reveals the challenges and practical significance of its research in scientific structure. Finally this paper also looks forward to relevant research directions that may be further expanded in the future.

Keywords: hybrid network community discovery clustering collaboration citation